NAMRL SPECIAL REPORT 93-6

ITERATED-BOOTSTRAP CONFIDENCE INTERVALS FOR THE MEAN

R.R. Stanny

Naval Aerospace Medical Research Laboratory
Naval Air Station
Pensacola, Florida 32508-5700

NAVAL AEROSPACE MEDICAL RESEARCH LABORATORY
51 HOVEY ROAD, PENSACOLA, FL 32508-1046

NAMRL SPECIAL REPORT 93-6

ITERATED-BOOTSTRAP CONFIDENCE INTERVALS FOR THE MEAN

R.R. Stanny

Approved for public release; distribution unlimited.

Reviewed and approved 20 Dec 93

A. J. MATECZUN, CAPT, MC USN
Commanding Officer

This research was sponsored by the Walter Reed Army Institute of Research under military interdepartmental purchase request No. 90MM0523. The project was monitored by the Naval Medical Research and Development Command under work unit 637.64A3M4637648995.AB-088.

The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, nor the U.S. Government.

Trade names of materials and/or products of commercial or nongovernment organizations are cited as needed for precision. These citations do not constitute official endorsement or approval of the use of such commercial materials and/or products.

ABSTRACT

Several reports indicate that nonparametric bootstrap confidence intervals (CIs) produced by the percentile method can yield overly liberal Type I error rates in small samples when the nominal α level is .05 or less. In the Monte Carlo simulations described here, percentile-method bootstrap 95% CIs for μ produced higher Type I error rates than standard parametric CIs in Gaussian and exponential samples of 40 or fewer observations. Iterated-bootstrap CIs for μ, however, yielded Type I error rates near α = .05 in Gaussian and exponential samples of as few as 10 observations. In exponential samples of 10 or more observations, iterated-bootstrap intervals controlled Type I errors more reliably than parametric intervals and were not obviously inferior to the parametric intervals when the data were Gaussian. Thus, ordinary percentile-method bootstrap CIs for μ may be of questionable accuracy when Type I error rates are to be controlled at values of α ≤ .05 or so. On the other hand, iterated-bootstrap CIs may be preferable to parametric CIs for data that come from a skewed distribution, such as the exponential, provided *n* is 10 or more. In samples of about 10 observations or more, iterated CIs may yield better Type I error control than parametric CIs when the data are skewed and nearly the same Type I error control when the data are Gaussian.

ACKNOWLEDGMENTS

I thank Michael Reddix, Andrew McCardie, John D'Andrea, and Kathleen Mayer for valuable critiques of earlier versions of this manuscript.

INTRODUCTION

Efron's bootstrap is a nonparametric technique for estimating variation in a statistic (Efron, 1979, 1982, 1988; Efron and Tibshirani, 1991). The method involves repeatedly drawing subsamples from an original data set. The statistic of interest is calculated in each subsample, and the frequency distribution of its values is taken as an approximation to the statistic's actual sampling distribution. The bootstrap is noted for generality and a remarkable ability to extract information from samples (Efron, 1982). Several investigators have noted, however, that standard bootstrap confidence intervals (CIs) for the correlation coefficient yield overly liberal Type I error rates in small samples when α is set to .05 or less (e.g., Efron, 1982; Rasmussen, 1987, 1988; Strube, 1988). This bias may derive from a tendency of the bootstrap to produce too few subsamples with extreme values of the statistic under examination (Young & Daniels, 1990). Indeed, Efron (1988) has observed that, although the bootstrap performs well with α set to .10, nonparametric bootstrap CIs perform better when "not pushed too far toward extreme coverage probabilities" (p. 295). However, in psychology, and many other areas of science, it is conventional to set Type I error rates to .05 or less.

Bootstrap resampling is performed by randomly drawing observations from an empirical data set. Observations are drawn with replacement and in such a way that each item in the original data set has an equal probability of entering a subsample. By convention, the number of observations drawn for each subsample, n, is usually set equal to the number of observations in the original sample. The number of bootstrap subsamples drawn, N, varies with the problem. Nonparametric bootstrap CIs are typically based on 500-2,000 subsamples; Efron (1988) suggests using a minimum of 1,000 subsamples.

A nonparametric, "percentile-method" bootstrap CI for an arbitrary statistic, θ, is generated by drawing N bootstrap subsamples, calculating the statistic's sample estimate, $\hat{\theta}$, in each subsample, finding the $100\alpha/2$ and $100(1 - \alpha/2)$ percentiles of the frequency distribution of the values of $\hat{\theta}$ thus produced, and taking the interval between these points as the range of a $100(1 - \alpha)\%$ CI. In this way, the percentile method "automatically" determines a set of approximate confidence limits associated with a given probability of Type I error. The percentile method does not require the assumption that the data follow any specific probability distribution. Its validity does, however, depend on the validity of assuming that distributions of bootstrap subsamples tend to reflect the forms of actual sampling distributions. The results of Rasmussen (1987, 1988) and Strube (1988) suggest that this assumption may sometimes be invalid when n is small and the desired α is .05 or less.

Several approaches to correcting the percentile method's bias have been proposed. The bias-corrected percentile method (Efron, 1982) yields CIs with better coverage properties than those of ordinary percentile-method CIs. However, the bias-corrected percentile method reintroduces parametric (Gaussian) assumptions. Furthermore, Monte Carlo studies have shown that the corrections it produces may be too small when n is small and α is set to .05 (two-tailed) or less (Strube, 1988). Another correction, the accelerated bias-corrected percentile method is evidently quite accurate in some situations (Efron, 1987). However, the accelerated bias-corrected method requires calculating an analytic correction factor that can be difficult or impossible to derive (Loh & Wu, 1987). For this reason, the accelerated bias-corrected method may be of questionable utility in routine data analyses conducted by nonspecialists.

The iterated bootstrap is a computationally intensive method of correcting the bootstrap's bias (Beran, 1987; Hall, 1986; Hall & Martin, 1988; Martin, 1990). Like the ordinary percentile-method bootstrap (and unlike the bias-corrected and accelerated bias-corrected methods) the iterated percentile-method bootstrap sets confidence limits automatically and requires no specific parametric assumptions. An iterated-bootstrap 95% CI for the mean can be calculated by first drawing N first-order bootstrap subsamples from an empirical sample. One then draws M bootstrap subsamples from each first-order subsample. This yields N sets of M second-order subsamples. The second-order subsamples drawn from each first-order subsample are used to calculate an ordinary percentile-method CI for μ. This produces N second-order, percentile-method CIs. The widths of the second-order CIs are then calibrated by adjusting their lower and upper cutoff percentages until values are found that cause 95% of the intervals to cover the sample mean. The empirically corrected cutoff percentages thus obtained are then substituted for 100α/2 and 100(1 - α/2)% in an ordinary percentile-method CI derived from the means of the first-order subsamples. The process of replacing the nominal cutoff percentages with the empirical percentages comprises the correction for the bootstrap's bias.

The Monte Carlo studies described in the following sections were carried out to examine the Type I error rates of iterated-bootstrap CIs for the mean in small samples from a decidedly nonGaussian population, the exponential. A CI for the mean expresses the precision with which a measurement has been obtained, and can be used to test hypotheses about the location of μ. For example, a two-sided, 100(1 - α)% CI for μ that fails to cover 0 can be used to justify rejecting the null hypothesis that $\mu = 0$ with significance α. Hence, the results described here are relevant to one-sample hypothesis tests, such as tests of differences between correlated observations and single-degree-of-freedom orthogonal polynomial contrasts. Exponential samples were chosen for examination because they can be generated fairly rapidly (a consideration in Monte Carlo studies of iterated bootstrapping) and because they represent a nearly worst-case scendario of sampling from a skewed distribution. Although psychological data that *precisely* follow an exponential distribution are probably rare, many types of data tend to exhibit the positively skewed form of the exponential (or a mirror image therof), particularly when floor or ceiling effects operate. Examples include accuracies on easy tests, error counts in reaction-time studies, answers on some rating scales and symptom questionnaires, and lapse probabilities during minor sleep deprivation.

METHODS

The simulations described here were written in Fortran-77 and run on an Intel 860 reduced instruction set computer installed in a desktop PC. Three types of CIs were examined: Gaussian-theory (Student's *t*), percentile-method bootstrap, and iterated percentile-method bootstrap. Random samples of data were drawn from two distributions, one Gaussian and one exponential. Gaussian samples were generated by randomly drawing values from a normal distribution with $\mu = 0$ and $\sigma = 1$. Exponential samples were generated by drawing values from an exponential distribution with $\mu = \sigma = 1$. Gaussian variables were generated by the direct method; exponential variables were generated by the inverse method (Zelen & Severo, 1970).

Gaussian-theory and percentile-method CIs were examined in Gaussian and exponentially distributed samples with sizes of 5, 10, 20, 40, 80, and 160. Iterated-bootstrap CIs were examined in Gaussian and exponentially distributed samples with sizes of 5, 10, 15, 20, and 25. One thousand confidence intervals were created in each experimental condition defined by a combination of CI type, probability distribution, and sample size. The observed Type I error rate in each experimental condition was calculated as the proportion of CIs that failed to cover μ.

Gaussian-theory CIs were calculated as $\bar{x} \pm t_{(1-\alpha/2)}(s_{\bar{x}})$, where $\bar{x}$ was the sample mean, $t_{(1-\alpha/2)}$ the critical value of Student's *t* corresponding to $1 - \alpha/2 = .975$ with $n - 1$ degrees of freedom, and $s_{\bar{x}}$ the sample estimate of the standard error of the mean. Percentile-method CIs were calculated by drawing $N = 1{,}000$ bootstrap subsamples from an empirical sample, calculating the means of the subsamples, sorting the N means, and taking 25th- and 975th-largest values as the 100α/2% and 100(1 - α/2)% limits of a 95% CI, respectively.

An iterated-bootstrap CI was calculated by drawing $N = 1{,}000$ first order bootstrap subsamples from an empirical sample, drawing $M = 1{,}000$ second order bootstrap samples from each first-order sample, and calculating the means of the second-order samples. The $N = 1{,}000$ cumulative frequency distributions of second-order subsample means produced by this method were searched for the percentile that exceeded the sample mean in 2.5% of the distributions of second-order means. A similar search was performed for the percentile that exceeded the sample mean in 97.5% of the distributions of second-order means. An ordinary percentile-method CI was then constructed from the means of the first-order subsamples after replacing the standard 100α/2 and 100(1 - α/2) percentage points with the corrected percentage points obtained from the search through the second-order means.

RESULTS AND DISCUSSION

Figure 1 illustrates the performance of the Gaussian-theory and percentile-method CIs in Gaussian samples. The Type I error rates of the Student's *t*-based intervals are near the nominal α level of .05 at all sample sizes. In contrast, the Type I error rates of the percentile-method intervals are much higher than .05 in small samples, averaging about .153 when *n* = 5 and not closely approaching .05 until *n* reaches about 40. In samples of about 40 or more observations, the percentile-method bootstrap works fairly well. In small samples, however, the percentile method yields confidence limits that are too narrow, thus yielding large numbers of Type I errors.

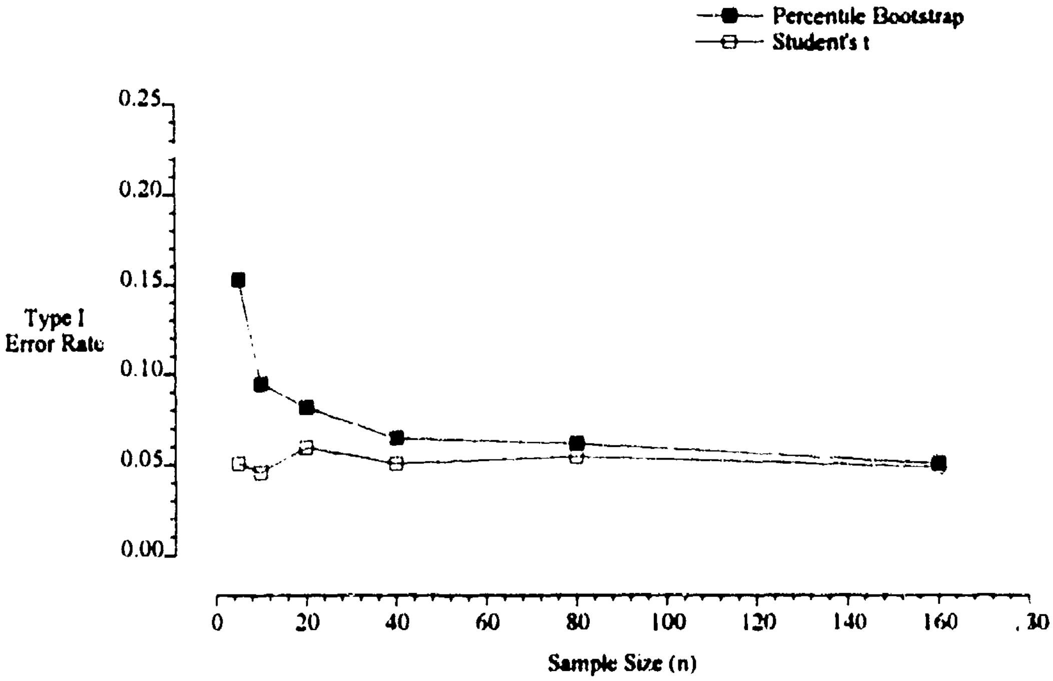

Figure 1. *Type I errors versus sample size for percentile-bootstrap and Student's-t intervals. Nominal Type I error rates were .05; samples were Gaussian.*

Figure 2 illustrates the performance of the Gaussian and percentile-method bootstrap intervals in samples drawn from the exponential distribution. Both types of intervals produce Type I errors at rates much higher than .05 in samples smaller than 20. Interestingly, the parametric intervals are less biased than the nonparametric intervals. Neither interval, however, performs especially well in exponential samples of the sizes examined here.

Figure 3 illustrates the performance of the iterated-bootstrap intervals in samples from Gaussian and exponential distributions. In samples of five observations, the iterated-bootstrap's Type I error rate is still very high, averaging about .138 in the Gaussian samples and about .134 in the exponential samples. In samples of size 10, however, the iterated-bootstrap's Type I error rate is near the nominal level of α = .05 in Gaussian samples and only slightly higher in exponential samples (averaging about .058 in the exponential data). In samples of 15 or more observations, the iterated bootstrap's observed Type I error rate is within sampling error of α = .05 in both

Gaussian and exponential samples. (By the normal approximation to the binomial, the standard error of the estimate of α = .05 in samples of 1,000 should be roughly $[(\alpha)(1 - \alpha)/1000]^{1/2} \doteq .007$.)

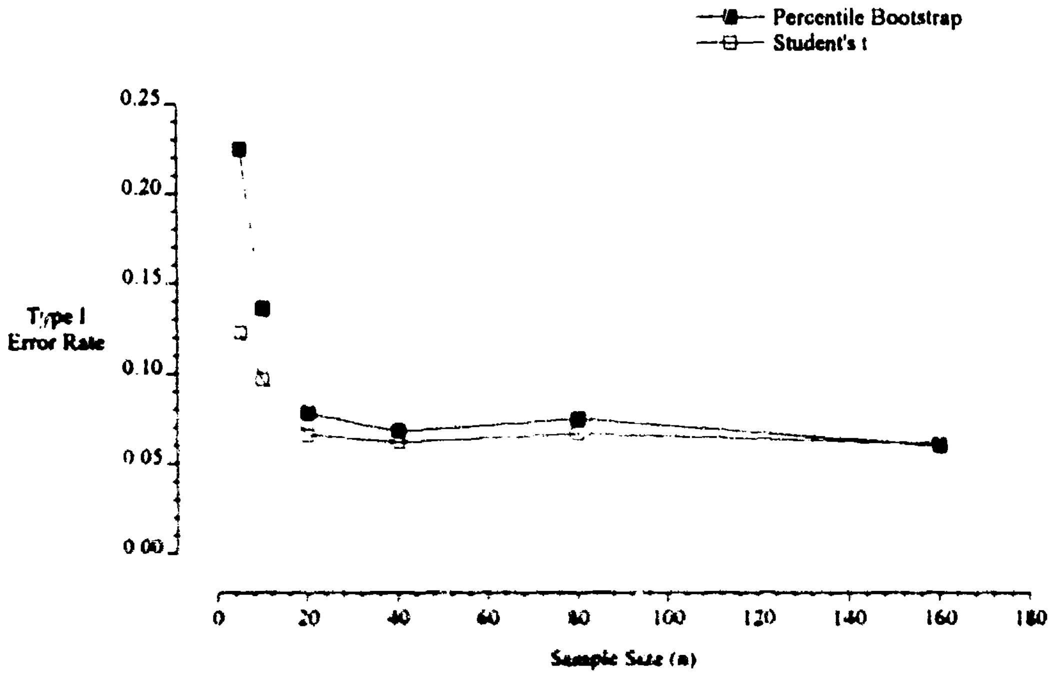

Figure 3 *Type I errors versus sample size for percentile-bootstrap and Student's-t intervals. Nominal Type I error rates were .05; samples were exponential.*

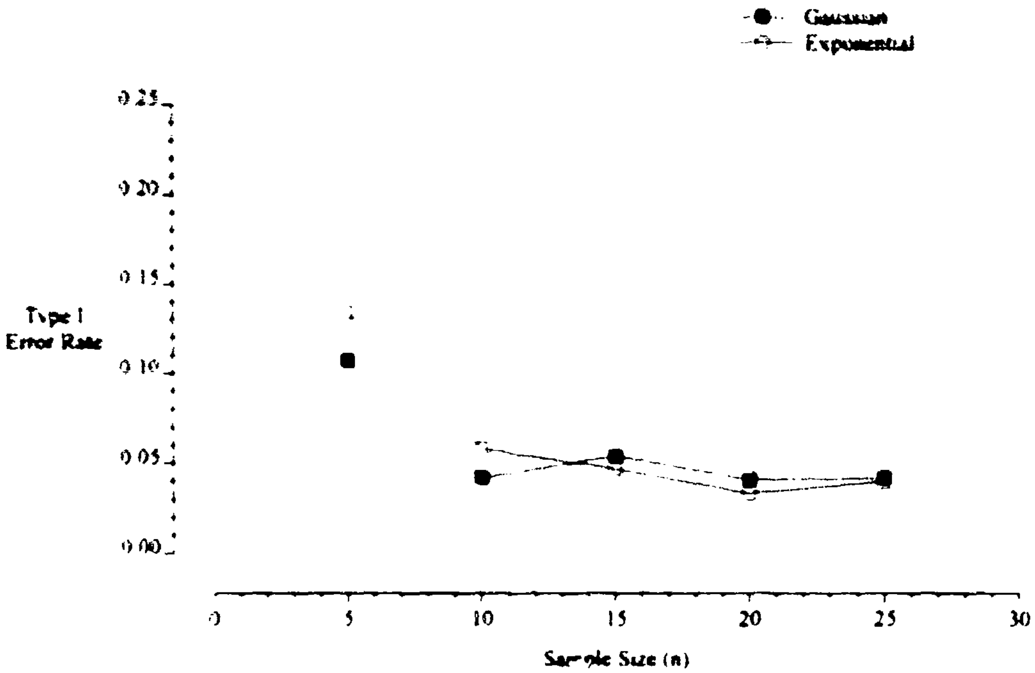

Figure 2 *Type I errors versus sample size for iterated bootstrap confidence intervals. Nominal Type I error rates were .05; samples were Gaussian and exponential.*

SUMMARY

None of the methods yielded Type I error rates near $\alpha = .05$ in samples of five observations drawn from an exponential distribution. Iterated-bootstrap intervals produced liberal Type I error rates in samples of fewer than about 10 observations. Student's *t*-based CIs were seriously biased in samples of 20 or fewer observations. Ordinary percentile-method bootstrap CIs were more biased than the Student's *t*-based CIs in samples this small.

Except in the smallest samples ($n = 5$), the iterated-bootstrap intervals yielded Type I error rates in Gaussian data that were indistinguishable from those of the Gaussian-theory intervals. Considering the fact that the *t*-based CIs are theoretically optimal when the data are Gaussian, this performance seems remarkable for a nonparametric technique. The failure of the iterated bootstrap in the $n = 5$ condition is disappointing but unsurprising given the uncertainties involved in reconstructing the sampling distribution of the mean from so few observations.

An iterated bootstrap consumes substantially more computer time than the ordinary bootstrap, which is itself computationally demanding. When N and M are set to 1,000, for example, an iterated bootstrap requires drawing $NM = 1{,}000{,}000$ subsamples, calculating 1,001,001 values of the statistic of interest, and sorting 1,001 arrays of 1,000 means. Some additional time is spent searching the arrays of second-order means for the adjusted percentile cutoffs. Calculations of this magnitude take time but are within the capabilities of reasonably fast desktop computers. A problem with $n = 15$ and $N = M = 1{,}000$, for example, might require about 6 min in efficiently coded Fortran on a 20-MHz 386-based machine with a math coprocessor.

RECOMMENDATIONS

1. Ordinary percentile-method bootstrap CIs for μ may be of questionable value when Type I error rates are to be controlled at values as low as .05. This is because, when $\alpha \leq .05$, percentile-method CIs may perform less well than parametric CIs in small samples and no better than parametric CIs in large samples.

2. When it is possible that data may have been drawn from a skewed distribution, like the exponential, iterated-bootstrap CIs may be preferable to parametric CIs if n is about 10 or more. Under these conditions, iterated CIs may yield better levels of Type I error control than parametric CIs when the data are skewed, and approximately the same level of Type I error control when the data are drawn from a Gaussian distribution.

REFERENCES

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika, 74*, 457-468.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171-185.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin, 104*, 293-296.

Efron, B., & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science, 253*, 390-395.

Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics, 14*, 1431-1452.

Hall, P., & Martin, M. (1988). On bootstrap resampling and iteration. *Biometrika, 75*, 661-671.

Loh, W.-Y., & Wu, C. F. J. (1987). Comment on "Bootstrap confidence intervals and bootstrap approximations." *Journal of the American Statistical Association, 82*, 188-190.

Martin, M. A. (1990). On bootstrap iteration for coverage correction in confidence intervals. *Journal of the American Statistical Association, 85*, 1105-1118.

Rasmussen, J. L. (1987). Estimating correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin, 101*, 136-139.

Rasmussen, J. L. (1988). "Bootstrap confidence intervals: Good or Bad": Comments of Efron (1988) and Strube (1988) and further evaluation. *Psychological Bulletin*, 104, 297-299.

Strube, M. J. (1988). Bootstrap Type I error rates for the correlation coefficient: An examination of alternate procedures. *Psychological Bulletin*, 104, 290-291.

Young, G. A., & Daniels, H. E. (1990). Bootstrap bias. *Biometrika*, 77, 179-185.

Zelen, M., & Severo, N. C. (1970). Probability functions. In M. Abramowitz and I. A. Stegun (Eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (pp. 925-995). Washington, DC: U.S. Government Printing Office.

Other Related NAMRL Publications

None.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE December 1993	3. REPORT TYPE AND DATES COVERED Interim 1991

4. TITLE AND SUBTITLE

Iterated-Bootstrap Confidence Intervals for the Mean

5. FUNDING NUMBERS

C 90MM0523
WU 63764A
3M4637648995.AB-088

6. AUTHOR(S)

R.R. Stanny

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NAVAEROMEDRSCHLAB
51 Hovey Road
Pensacola, FL 32508-1046

8. PERFORMING ORGANIZATION REPORT NUMBER

NAMRL Special Report 93-6

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

NAVMEDRSCHDEVCOM
National Naval Medical Center
8901 Wisconsin Avenue
Bethesda, MD 20889-5606

Walter Reed Army Institute of Research
Washington, DC 20307-5100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Several reports indicate that nonparametric bootstrap confidence intervals (CIs) produced by the percentile method can yield overly liberal Type I error rates in small samples when the nominal α level is .05 or less. In the Monte Carlo simulations described here, percentile-method bootstrap 95% CIs for μ produced higher Type I error rates than standard parametric CIs in Gaussian and exponential samples of 40 or fewer observations. Iterated-bootstrap CIs for μ, however, yielded Type I error rates near $\alpha = .05$ in Gaussian and exponential samples of as few as 10 observations. In exponential samples of 10 or more observations, iterated-bootstrap intervals controlled Type I errors more reliably than parametric intervals and were not obviously inferior to the parametric intervals when the data were Gaussian. Thus, ordinary percentile-method bootstrap CIs for μ may be of questionable accuracy when Type I error rates are to be controlled at values of $\alpha \leq .05$ or so. On the other hand, iterated-bootstrap CIs may be preferable to parametric CIs for data that come from a skewed distribution, such as the exponential, provided n is 10 or more. In samples of about 10 observations or more, iterated CIs may yield better Type I error control than parametric CIs when the data are skewed and nearly the same Type I error control when the data are Gaussian.

14. SUBJECT TERMS

Statistics, Nonparametric Statistics, Bootstrap, Confidence Intervals

15. NUMBER OF PAGES
13

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT SAR